Bitirme Projesi Raporu

Hilal IŞIK-141180034
Saliha ERCİYAS- 131180028

BM 496 Bilgisayar Projesi II
Akıllı ve Teknolojik Ürünler

MAYIS 2018

1.Özet

Bu rapor kapsamında gerçekleştirdiğimiz İngilizce konuşma tanıma sisteminin yapısından bahsedilecektir. Öncelikle, konuşma tanıma ile ilgili literatur taramasına yer verilecektir. Daha sonra ise konuşma tanıma sisteminin gerçekleştirilen adımları üzerinde durulacaktır. Sistem üzerinde gerçekleştirdiğimiz başarılı ve başarısız deneylere yer verilecektir. Sistemde kullanılan sınırlı dağarcıklı ve geniş dağarcıklı deneyler üzerinde durulacaktır. Modern saklı Markov modeli tabanlı konuşma tanıma sistemlerinin İngilizce için değişik sınama koşullarındaki başarımı gösterilecektir. Basit testlerde hata oranı %2 civarında olurken, geniş dağarcıklı modellerdeki hata oranının fazla olmasının nedeni üzerinde durulucaktır. Bu nedenler raporun, tartışma kısmında ayrıca ele alınacaktır. Yaptığımız çalışmanın, ingilizce konuşma tanıma sistemi konusunda yapılacak çalışmalara temel teşkil ettiği düşünülmektedir.Ayrıca bize de doğal dil işleme ve yapay zeka konusunda zor ama zevkli bir tecrübe yaşatmıştır.

Raporun birinci bölümünde giriş yapılarak, genel olarak konuşma tanıma sisteminin avantajlarından ve karşılaştığı zorluklardan bahsedilecektir. İkinci bölümünde genel olarak konuşma tanıma sistemlerinin sahip olduğu yapı üzerinde durularak, kullandıkları yöntemlerden ve özelliklerinden bahsedilecektir. Üçüncü bölümde, sistemin eğitilmesi için kullandığımız verinin yapısını anlatacağız. Dördüncü bölüme bizim deneylerimizi yaparken kullandığımız sistemi açıklayarak başlayacağız. Bunun için konuşma tanıma sistemi özelliklerinin bizim uygulamamızda nasıl olduğunu açıklayacağız.

2.Giriş

Konuşma tanıma, insan sesinin bilgisayar tarafından algılanmasıdır. Diğer bir ifadeyle, konuşma verisini yazıya çeviren bir yazılım olarak da düşünülebilmektedir. Bu tür sistemlerin ve prpjelerin gerçekleştirilmesi, insam makine etkileşimi daha kolay ve verimli bir hale getirecektir. Ancak bu sistemlein geliştirlimesi birçok zorlu sürec içide barındırmaktadır. Her dil için farklı ses tanıma yazılımları oluşturulmak zorundadır, dolayısıyla her farklı dilde ses tanıma yazılımı o dile has özelliklerden kaynaklanan sorunları çözmek zorundadır.

Bütün bu zorluklara rağmen, konuşma tanıma sistemleri büyük avantajlar sağlamaktadır. Bu avantajlardan aşağıda paragrafar halinde kıaca bahsedilmektedir.

Konuşma tanıma sistemleri metin dikte ettirmek için de kullanılabilir. Herhangi bir editör programına sesle yazı dikte ettirmek, bu iş için harcanan süre ve emeği azaltacaktır. Ek olarak, doğrudan ses dosyalarından yazıya geçiş yaptıralabilir ve duyma güçlüğü çeken kişilerin yayınları takip etmesine olanak sağlar.

Ses tanıma sistemleri, gelişmiş çeviri yazılımlarıyla birlikte kullanılarak da faydalı olabilir.

Birleştirilmiş bir sistemle farklı ana dile sahip iki kişi ortak bir dil bilmek zorunluluğu olmadan iletişim kurabilirler ve bu türden bir aygıt uluslararası konferanslarda simultane çeviri yaparak faydalı olabilir.

Gelişmiş bir ses tanıma sistemi, makinelere ses ile komut vermeye olanak sağlar. Teknolojik gelişmelerle televizyonlar, arabalar ve diğer elektronik aletler hiçbir aracı alet olmadan sesle yönetilebilir.

Konuşma tanıma programları askeri açıdan da çok önemlidir. Hava Kuvvetlerinde konuşma tanıma, pilot iş yükünü azaltmak için belirli bir potansiyele sahiptir. Hava Kuvvetleri'nin yanı sıra bu tür programlar helikopterlerde kullanılmak üzere eğitilebilmektedir.

Birçok şirket çağrı merkezlerinde görevliler bulundurmaktadırlar ve bu da para ve insan gücü kaybına sebep olmaktadır. Bu kaybın önüne çağrı merkezindeki görevlilerin işini yapacak bir ses tanıma sistemi ile geçilebilir. Bu şekilde harcanan süre ve emek de azalacaktır.

Ses tanıma sistemleri, gelişmiş çeviri yazılımlarıyla birlikte kullanılarak da faydalı olabilir. Birleştirilmiş bir sistemle farklı ana dile sahip iki kişi ortak bir dil bilmek zorunluluğu olmadan iletişim kurabilirler ve bu türden bir aygıt uluslararası konferanslarda simultane çeviri yaparak faydalı olabilir.

3. Konuşma Tanıma Probleminin Süreci

Temel olarak konuşma tanıma sisteminin yaptığı görev bir konuşma verisini almak ve ne söylenildiğini tahmin etmektir. Yani sistemin girdisi konuşma verisidir ve çıktısı da tahmin edilen cümledir (hipotez cümlesi).Sistemin tahmin yapabilmesi için iki ana bölüme ihtiyacı vardır: öznitelik çıkartıcı ve dil çözümleyici. ***Öznitelik çıkartıcı,*** gelen ses dalgayapısındaki veriyi akustik öznitelik dizilerine dönüştürmektedir. Bu şekilde dil çözümleyicisinin kullanacağı bir yapı oluşturmaktadır. Daha sonra, dil çözümleyici bu öznitelikleri çıkartılmış diziyi alır ve cümleyi tahmin etmeye çalışır. Tahmin ederken iki temel model kullanılır: ***Dil modeli ve akustik model***. Bu modeller hakkında da detaylı bilgi aşağıda ayrıntılı bir şekilde verilecektir.

3.1. Speech Recognition Temel Adımları

- ***Ses Girişi:*** İlk olarak sesin alınabilmesi için mikrofondan yardım alınmıştır. Mikrofon ile ses girişi yapılmaktadır. Alınan ses, pc ses kartında digital bir şekilde gösterilmektedir.
- ***Sayısallaştırma:*** Alınan ses analog sinyal halindedir. Analog sinyalin, sayısal sinyale dönüştürülmesi gerekmektedir. Bu yüzden projemizde ikinci adım olarak, sayısallaştırma yapılmaktadır. Proje de dikkat ettiğimiz nokta ise, sürekli sinyalin ayrık sinyale dönüştürürken, kuantizyon sürecidir. Kuantizyon süreci, sürekli bir değer aralığına yaklaşma sürecidir.

- ***Akustik Model:*** Sesli konuşma kayıtları ve metin transkripsiyonları alınarak ve her kelimeyi oluşturan seslerin istatistiksel temsillerini oluşturmak için yazılım kullanılarak bir akustik model oluşturulmaktadır.. Konuşmayı tanımak için bir konuşma tanıma motoru tarafından kullanılır. Yazılım akustik modeli kelimeleri fonemlere ayırmaktadır.
- ***Dil Modeli :*** Dil modellemesi konuşma tanıma gibi bir çok doğal dil işleme uygulamasında kullanılır. Bir dilin özelliklerini yakalamaya çalışır ve konuşma sırasındaki bir sonraki sözcüğü tahmin etmeye çalışır. Yazılım dili modeli, fonemleri yerleşik sözlükteki kelimelerle karşılaştırır.
- ***Konuşma Motoru:*** Konuşma tanıma motorunun görevi, giriş sesini metine dönüştürmektir . Bunu başarmak için her türlü veriyi, yazılım algoritmalarını ve istatistiklerini kullanır. İlk operasyonu daha önce tartışıldığı gibi dijitalleştirmedir, yani daha fazla işlem için uygun bir formata dönüştürülür. Ses sinyali uygun formatta olduğunda, en iyi eşleşmeyi arar. Bunu bildiği kelimeleri göz önünde bulundurarak yapar, sinyal tanındığında ilgili metin dizesini döndürür.
- ***Konuşma Tanıma Programlarının Kullanımları :*** Temel olarak konuşma tanıma iki temel amaç için kullanılır. Konuşma tanıma bağlamında yer alan ilk ve en önemli dikte, konuşulan sözcüklerin metne çevrilmesi ve ikinci olarak bilgisayarın kontrol edilmesi, yani muhtemelen bir kullanıcının farklı uygulamaları sesle çalıştırması için yetkin olacak bir yazılım geliştirmektir.

  Temel olarak tahmin işlemi aşağıdaki Bayes denklemi sayesinde gerçekleştirilmektedir. Teorik bilgi vermesi açısından, aşağıda bayes denklemi yer almaktadır.

  $$\hat{W} = \arg\max_{W} P(W|A) = \arg\max_{W} P(A|W)P(W) \qquad (1)$$

  Bayes denkleminden kısaca bahsetmemiz gerekirse, Wˆ hipotez cümlesidir. Bu cümle eşitliğin ikinci tarafında da görülebileceği gibi olası cümleler içinden akustik diziyle uyuşma ihtimali en yüksek olan cümledir. P(W|A) verilen akustik veriye göre (A) mevcut cümlenin (W) akustik diziyle eşleşme olasılığını simgeler. Eşitliğin ikinci kısmı doğrudan Bayes kuralının uygulanması ile elde edilir. Burada P(W), W cümlesinin oluşma olasılığıdır ve dil modeline göre hesaplanır. P(A|W) verilen cümleye göre bizdeki akustik dizinin oluşması olasılığıdır ki bu da akustik modelle hesaplanır. Bayes kuralından gelen P(A) W'dan bağımsız olduğu için sadeleştirilmiştir [4-5].

3.1.1. Öznitelik Çıkarımı

Konuşma tanıma sisteminin tasarımında yapılan önemli bir varsayımdır. Bir konuşma sinyalinin segmenti bir birkaç milisaniye aralığıdır. Bu nedenle konuşma sinyali kare olarak adlandırılan bloklara ayrılmaktadır. Bir karenin boyutu yaklaşık 25 m-sn'dir. İki blok arasındaki mesafe ise yaklaşık olarak, 10m-sn'dir. Burdan da anlaşılacağı üzere, konuşmayı analiz edeceğimiz yerler çerceve boyutlarıdır.

Peki konuşma sinyallerini kullanabilmek için hangi parametreleri kullanmamız gerekmektedir? Bu soru ise, üzerinde durduğumuz ikinci başlık olmaktadır. Okuduğumuz makalelerden çıkardıpımız sonuç, geçmişte ***Doğrusal Katmin Katsayılarının*** ilk kullanılan parametre yöntemleri olduğu yönündeydi. Ancak bu yöntem ile yüksek performans elde edilemediğini öğrenince, bu parameter yöntemini bıraktık. Daha sonra ise, son zamanlarda sık kullanıla ve bizim birinci dönemki bitirme projemizde de yoğun bir şekilde anlamaya çalıştığımız, ***Line Spektral Frekansları*** ve ***MelCepstral Katsayıları*** üzerinde durduk. Ancak bu yöntemlerle elde ettiğimiz başarının sonuçları yüksek olmadı. Zamanın da hızla geçtiğini düşünerek, ***Fourier dönüşümüne*** geçtik. Eşit frekans aralıklı üçgen filtreler oluşturulur ve ses dalgasının Fourier dönüşümü bu filtrelerle çarpılarak toplanır. Böylelikle filtrelerdeki değerlere göre ağırlıklı bir toplam elde edilir. Yeni oluşturulan bu veriler ve ***MFCC (Mel-frequency cepstral coefficients)*** kullanılarak vektör öznitelikleri çıkartılır.

3.1.2. Akustik Model

Konuşma tanıma sistemlerinde akustik model denildiğiinde akla gelen ilk yöntem HMM (Hidden Markov Modeli)'dir. Yukarıda akustik modelin fonemlere ayrılması üzerinde durulmuştur. Fonem bazı modellerde 3 adet Hidden Markov Modelinden oluşmaktadır. HMM'in model parametreleri durum değiştirme olasılıkları (state transition probabilities - aij) durum gözlenme yoğunlukları (state observation density - bj(o)) ve başlangıçtaki durum dağılımlarıdır . bj(o) 'nin hesaplanması Gauss karışım dağılım olasılık fonksiyonlarının kullanılmasıyla hesaplanır. Kullanılabilecek Gauss karışımı sayısı artırılabilir [5].

Öte yanda dil modelleri, eğitim metinlerinden eğitilir ve cümlelerin o dildeki olasılıklarını oluşturur. Eğer cümlelerin kelimelerden oluştuğunu varsayarsak, wi kelimelerinden oluşan W cümlesinin tanınma olasılığı aşağıdaki gibidir:

$$P(W) = P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, ..., w_{i-1}) \quad (2)$$

Dil modelleri uygulamalarında genellikle N-gram modeller kullanılır, bu modeller de denklemin sağ tarafında sadece w_i'den önceki N-1 kelimeyi kullanarak denkleme yaklaşır. Böylece yukarıdaki denklem aşağıdaki denkleme dönüşmüş olur:

$$P(W) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-N+1}, ..., w_{i-1}) \quad (3)$$

Bu N-gram olasılıkları ise aşağıdaki formülle hesaplanır:

$$P(w_i \mid w_{i-N+1},\ldots,w_{i-1}) = \frac{C(w_{i-N+1},\ldots,w_{i-1},w_i)}{C(w_{i-N+1},\ldots,w_{i-1})} \quad (4)$$

Buradaki $C(w_{i-N+1},\ldots, w_{i-1}, w_i)$ belirtilen N kelimenin eğitim metinlerindeki geçme sayısını gösterirken, $C(w_{i-N+1},\ldots, w_{i-1})$ ise belirtilen N-1 kelimenin metinlerde geçme sayısını gösterir. N=2 durumunda ise bigram dil modeli elde edilir [6].

Önemli bir nokta ise, smoothing yöntemidir. Bu yöntemin amacı, eğitim verisinde görülmeyen kelimelerin tanınmalarında uygulanmasıdır. Tanımlanan bu uygulamalara birer olasılık değeri atanmaktadır. Bu dil modellerinde oluşturulan istatistiksel dağılımlara göre kelimelerin tanınma olasılıkları, aynı akustik modelde olduğu gibi, değişim gösterir.

3.2. Değerlendirme Ölçütü

Konuşma tanıma sisteminin performansını ölçmek için, en büyük öğrendiğimiz bilgi, monofon tanıma sisteminin etkili olmadığıdır. Çünkü bir fonem birden fazla sese karşılık gelebileceği için monofon tanıma sistemi etkili olmamaktadır. Bu yüzden trifonlara ihtyiaç duyulmaktadır. Peki trifon nedir? Trifonlar her fonemin hem sağında hem solunda bulunmaktadır.Bu sebeple trifonlar kullanmak akustik değişkenliği düşürür ve daha etkili bir tanıma sağlar

4.Veritabanı

Akustik modellerin eğitiminde temel olarak SUVoice ve Metu1.0 veritabanları örneklerinden yararlanılmak istenildi.

4.1.SUVoice Veritabanı

Sabancı Üniversitesi öğrencileri tarafından bir proje dersi kapsamında her dönem düzenli olarak toplanan, 6 yıllık bir konuşma veritabanıdır. Toplam 3444 okuyucunun ses kayıtlarından oluşturulmuştur. Yaklaşık 70 saati sessizlik olan 185 saatlik konuşma verisi içerir. 46 farklı metin dosyasından 4087 özgün cümlenin okunmasıyla oluşturulmuştur. Konuşmacıların 2055'i erkek, 1389'u ise bayandır.

4.2. METU 1.0 Veritabanı

Orta Doğu Teknik Üniversitesi ve University of Colorado at Boulder'ın ortak çalışması olan SONIC konuşma tanıma sisteminin Türkçe'ye uyarlanması projesi için ODTÜ'de toplanmış verilerden oluşmaktadır. Yaklaşık 500 dakikalık ses verisi mevcuttur. Her konuşmacı yaklaşık 40 cümle okumuştur ve 2462 özgün cümleden oluşmaktadır. Veriler 68'i erkek 52'si kadın 120 kişinin ses kayıtlarını içerir [3].

5.Deneyler ve Tartışma

Gerçekleştirilen programın eğitilmesi için, eğitim verisi ve test verisi olmak üzere ikiye ayrıldı.

5.1. Konuşma Tanıyıcı Yapısı

Konuşma tanıyıcı sistemin eğitimi için temel olarak HMM kullanılmaktadır. Öznitelik çıkarımları da MFCC yöntemine göre yapıldı. Akustik modelin eğitimi için İngilizce'deki 26 harfe karşılık gelen 26 fonem kullanıldı. Bunlara ilaveten kelimeler arasındaki kısa boşlukları yakalamak için kısa durak ve cümle aralarındaki boşlukları yakalamak için sessizlik modelleri eğitildi. Her bir model için 3 durum oluşturuldu ve her bir durum için 12 Gauss karışımı kullanıldı. Öncelikle bu 34 fonem üzerinden monofon bir eğitim gerçekleştirildi. Daha sonra mevcut fonemler kullanılarak trifonlar oluşturuldu. Bütün 3'lü fonem kombinasyonlarını eğitmeye yetecek kadar veri olmadığı düşünüldüğünden bağlı- durum (tied-state) trifonlar kullanıldı

Dil modeli olarak ise Osman Büyük'ün "*Sub-Word Language Modeling for Turkish Speech Recognition"* isimli yüksek lisans tezinde kullandığı kelime tabanlı bi-gram dil modelini kullanıldı.

5.2.Test Verisi

Test deneyleri, sınırlı ve geniş dağarcıklı olmak üzere iki ayrı kategoride yapılmıştır. Deneylerde kullanılan kategoriler, SUVoice veri tabanının kullanılmayan alt kümeleri oldu.

5.3 Deney Sonuçları

Şu an raporumuzda deney sonuçlarına yer verilmemektedir.

Hala deney sonuçlarını yükseltme aşamasında uğraşılmaktadır.

5.4. Tartışma

Yapılan araştırmalar ve çalışmalar doğrultusunda başarımın zamana bağlı olarak değişkenlik gösterdiği saptanmıştır. Bunun en büyük nedeninin ise sesin, bir seslendirilişi ile bir diğer seslendirilişi arasında, aynı kişi seslendirse dahi çokça farklılık göstermesi olarak görülmüştür. Diğer bir nedenin ise, ortam gürültüsü olduğu çalışmalar sırasında tespit edilmiştir.

6.SONUÇ

Bu çalışma sonucunda, mevcut tekniklerden yararlanmak suretiyle ses tanıma sistemi oluşturulmuş ve elde edilen bu sistem, bu konuda çalışma yapacak kişilere bu tekniklerin uygulama içinde kullanımı ve ilişkilendirilmesi hakkında örnek oluşturacaktır. Ayrıca bu ve benzeri uygulamalar sayesinde de, insanlar listening speaking becerilerini geliştirebileceklerdir. Çünkü böyle bir uygulamaya hem kolayca erişebilecek, hem de hiçbir maliyet ödemeyeceklerdir.

7.KAYNAKLAR

[1] Jurafsky, D., and Martin, J. H. (2000) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (1st Ed.), Prentice Hall, New Jersey, 934s.

[2] Kale, K. R. (2002) Dynamic Time Warping, http://www.cnel.ufl.edu/~kkale/dtw.html (04.04.2002)

[3] Büyük, O., "*Sub-word language modeling for Turkish speech recognition,*" M.S. thesis, Sabanci University, Istanbul, Turkey, 2005.

[4] Erdogan, H., Buyuk, O., Oflazer, K., "Incorporating language constraints in sub-word based speech recognition," presented at *IEEE Automatic Speech Recognition and Understanding Workshop*, Cancun Mexico, 2005.

[5] Young, S., et al., "*The HTK Book (for HTK Version 3.4),*" Cambridge University Engineering Department, 2006.

[6] Salor, Ö., Pellom, B., Çiloğlu, T., et al, "On developing new text and audio corpora and speech recognition tools for the Turkish language," presented at *7th International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002.

[7] Jurafsky, D., and Martin, J. H. (2000) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (1st Ed.), Prentice Hall, New Jersey, 934s.

[8] Kale, K. R. (2002) Dynamic Time Warping, http://www.cnel.ufl.edu/~kkale/dtw.html.

[9] Smith, J. O. (2003) Mathematics of The Discrete Fourier Transform with Audio Applications, http://ccrma.stanford.edu/~jos/mdft/Alias_Operator.html

[10] Cook, S. (2002) Speech Recognition HOWTO, http://www.gear21.com/speech/index.html.

[11] Ibarra, O. M., and Curatelli, F. (2000) A Brief Introduction to Speech Analysis and Recognition, An Internet Tutorial., http://www.mor.itesm.mx/~omayora/Tutorial/tutorial.html.